NOTES ON YEAR 2 KICKOFF MEETING AND PLANS:

Morning meeting – Year 2 experiment planning:

Needs:

- Present dynamic stimuli
- Increase sample size
- Relate objective measure of visual properties of HUDs to pilot perceptions of clutter.
- Assess pilot performance with HUDs at various stages of approach.
    - Validate subjective rating scales for assessing clutter, based on pilot performance ("redundant/orthogonal", "monochromatic/colorful", "not salient/salient", "sparse/dense", and "monotonous/variable").
- Relate level of perceived clutter to shifts in pilot attention from HUD to HDDs.

Objective:

- Identify physical display features significant in predicting pilot perceptions of clutter (create multidimensional, objective model of clutter).
- Identify perceived clutter "threshold" leading to attention shifts or performance degradations.
- Prototype and assess validity of multidimensional subjective measure of clutter.

Method:

- Select 3-6 E/SVS HUD configurations for evaluation based on MDPREF analysis from Year 1.
    - Displays must be representative of "low", "medium" and "high" clutter conditions (1-2 displays to represent each subjective level).
    - Characterize all displays in terms of measurable visual properties.
    - Need to select properties based on access to and ease of measurement (e.g., density, display groups, target-to-background contrast ratio, luminance, colors (hue saturations, etc.)
    - Develop subjective rating scales on perceptual dimensions of clutter.
        - Using multiple scales.
    - Use pairwise comparisons for clutter dimensions, based on training experience with HUDs, to establish rankings.
    - Use dimension rankings and ratings at end of each test trial to compute overall clutter score (rank-weighted sum of ratings across dimensions).
- Run experiments to test pilot performance with of "low", "medium" and "high" clutter HUDs and collect subjective ratings.
    - Initially discussed using video "snip-its" of phases of flight during approach and conducting verbal protocol with pilots as they fake actions at simulated FCP interface.
        - Videos would be of initial descent and on final.

Performance measures would be limited to decisions on actions and timing of actions (could not capture actual aircraft control and flight technical errors).
(Simulator staff said button camera recordings of HUD and out-of-cockpit view in simulator would be no good for creating videos.
Also said it would be possible to create videos using direct output from simulator graphics boards.)

Also discussed using X-plane simulator and integrating imagery of E/SVS HUD.
NC State has X-plane simulator with SVS development kit, can model commercial platforms (747, 757)
Simulator provides lower level of realism of simulator – soft FCP and CDU.
Importing E/SVS images in X-plane may be time consuming.
Synchronizing a separate E/SVS display with X-plane monitors would also be time consuming.

IFD simulator use was not what was proposed for Year 2 and Lance was not certain resources would be available.

Procedure - One display would be tested in each trial with trials lasting approximately 15-20 min.

Clutter of display would be rated for each approach segment.
Pilots would be asked to designate display as primary reference or identify when they would want to "declutter" the HUD.
Another proposal to identify a clutter threshold was to look for fall-offs in performance depending upon display conditions.
(This would require continuous performance measures with high sensitivity to HUD content.)
Eye-tracking was also proposed as a method to assess pilot gaze pattern depending upon the content of the HUD.
If the HUD is "cluttered", do pilots still look at it or go heads-down?

Data analysis:

Need identify visual properties of displays that are significant in explaining perceived clutter (ANOVA models may be sufficient).
Also need to conduct logistic regression analysis on pilot attentional responses of aircraft control measures to identify level of perceived clutter that appears to be trigger for change in direction of attention or in control behaviors to main an accurate flight path.

Visit to simulator:

We will be using the CaD-CE setup for our Year 2 experiment.

Hand-flying could involve curved-approach into KRNO starting from Mustang.
Can also use ILS approach with autoflight modes.

CaD-CE Setup can support video development or actual simulator exercises.

To be consistent with Year 1 study, we would use wireframe SVS in all trials.

This will need to be implemented by simulator staff.

Fusion of SVS and EVS imagery in specific display conditions can be adjusted at co-pilot controls.

TCAS symbology (diamonds and plus, minus indicators) can be added to displays.

Display concepts to be investigated…

Wireframe SVS used in Y1, but IFD currently outfitted with photorealistic imagery → should not be difficult to implement wireframe

NASA IFD fuses transition of SVS (above 500ft) to EVS (below 500ft)

IFD does not currently include TCAS symbology → should not be difficult to implement traffic icons

For Y2 study, small set of display configurations (6 to 9), representing low, medium, high clutter will be implemented; definitions will come from those used in Y1 study

Briefing:

Points of clarification on Year 1 research:

Need to present pilots with definitions of terms used as anchors to subjective rating scales on underlying dimensions of clutter.

Each pilot may have different internal definition compromising consistency of ratings and identification of important dimensions.

(Lance, commented that there was convergence of pilot opinion; however, we probably could demonstrate some convergence with any data set depending upon the methods we used.)

Need to be clear in any reports that one 4 hr. session was conducted with each pilot including 4 test trials.

There were variations in weather and field conditions across trials (not "runs" of a simulator)

(There were no questions or concerns posed with respect to the various statistical analyses and the reported results.

Reported that pilots interpreted the "unsafe/safe" pair as outcome of clutter and not underlying display quality driving perceived clutter. Said that pair would not be used as basis for subjective ratings scale in Year 2.

Additional rating subscale to be developed based on principal component from FA on "dynamic nature of displays.")

Critical comments on Year 1 research:

Mike Monroe (sp?) commented that he did not think pilot designation of the HUD as primary or secondary flight reference (at the close of trials) made sense, as pilots are required (or told) to use the PFD and ND, among other displays. Therefore the question

may not be a sensitive indicator of HUD conditions that lead to shifts in attention or decrements in performance.

In the Year 1 study, we defined clutter threshold based on "preference"

In Y1, used "would you use as primary display?"

Mike said procedural rules dictate use as primary display, so would not be sensitive to display clutter, content changes – this is what we found in Y1

Mike and Steve suggested use of "would you declutter display?"

(Said much thought needs to go into the wording of this, and the criteria used.)

Lance suggested asking if pilots would want to use "declutter" option at close of trials. This might be more of an indicator of pilot preference for HUD design than an indicator of a clutter-threshold leading to a performance decrement.

[Clutter threshold based on performance]

Emily/Amy suggested looking at performance measures to identify decrements relative to HUD configurations and clutter ratings.

(Emily and Amy would like to see the use of an objective "performance threshold" where we can identify the performance drop-off as a function of increased clutter.)

One issue here is that a near-continuous measure of control behavior that is highly dependent on perception may be necessary to detect performance problems.

Flight technical error was suggested but the acceptable range of control for identifying errors is often so broad that this may not be a sensitive indicator of decrements.

Other issues identified by Emily and Amy include:

What measures should we use? Potential options:

Flight technical error

Control inputs

Visual attention (eye-tracking, see below)

SA probes (or even more granular SA1 vs. SA2 vs. SA3)

Procedural errors

What do we do when measures conflict? For example, we see a drop-off in one measure but not another. How do we prioritize these measures?

What do we do if we see a gradual drop-off? What if the slope is equivalent across all levels of clutter? How can we say when performance is now "too poor?"

Consider using pilot technical standards to define performance drop-offs

There was some additional discussion of eye-tracking and the drawback of laborious data handling and analysis. (There is a need to define measures that can be collected in a reasonable amount of time and that do not require levels of analysis beyond the effort proposed for the grant.)

Mike Monroe also commented that a limitation of the Year 1 study was that a range of workload conditions, representative of actual piloting, was not investigated.

He said it is critical to know when display clutter becomes an issue under high workload conditions. (This is related to the Year 2 study recommendations below.)

Mike discussed the importance of contextual factors and cognitive workload in contributing to perceptions of clutter

Consider how these variables may be encoded in the model

Collect NASA-TLX ratings?

Suggestions for Year 2 research:

Steve Young said that we need to make clear what we mean by dynamics; that is, we are referring to display dynamics or movement of non-iconic and iconic information in HUD.

Afternoon meeting - Discussion on Year 2 experiment:

Needs:

There will need to be consultation with NASA Langley on those HUD configurations to be investigated in the experiment.

NCSU will provide a list of multiple configurations representing "low", "medium" and "high" clutter. NASA researchers will be asked to select among these configurations to set the experimental conditions. (There should be some basis for this selection - i.e., we selected those configurations that are mostly likely to appear in the cockpit in the near future or to be used by pilots.

Images of all display configurations to be analyzed in Year 2 will need to be provided to NCSU in advance of experimentation for visual property analysis.

Low-level display feature measures of interest (will discuss further with Karl):

Luminance

Contrast (iconic vs. non-iconic symbology)

Density (global vs. local)

Saturation

Grouping

Occlusion

Subjective response measures:

The approach to the development of the multidimensional subjective measure of clutter will be based on the design of the NASA-TLX, multidimensional measure of cognitive workload.

Visual analog scales of measured length are proposed for the ratings on the underlying dimensions of clutter.

On the subjective rating scales, Emily and Amy suggested a careful mapping of selected semantic pairs to principal components (PCs) as well as low-level display features

Redundant/orthogonal

Relates to "similarity of features" PC

What low-level display feature(s) does this pair map to?

Monochromatic/colorful

Relates to "clarity of features" PC

What low-level display feature(s) does this pair map to?

Not salient/salient

Relates to "global display density" PC

What low-level display feature(s) does this pair map to?

Sparse/dense

Relates to "global display density" PC

What low-level display feature(s) does this pair map to?

Note that we have not selected a semantic pair related to the "dynamic nature of displays" PC

Where did monotonous/variable, static/dynamic, and grouped/ungrouped fall in the transfer function analysis?

Emily and Amy suggest retaining one of these pairs for subjective rating scale development

If one of these pairs is retained, what low-level display feature(s) does this pair map to?

Test approach:

Using video –

Steve Young thought video might be a good approach in order to take advantage of pilots at distributed locations.

Steve mentioned that Dryden has 20 pilots available for experiment participation

As mentioned above, could use video snippets if IFD simulator is not available (though Randy, Steve, and Mike seemed against this)

Would get verbal protocol

Would replicate automated approach

Would measure timeliness of action relative to information in the display

Mike Monroe had concerns with the approach in that we would not be able to observe and analyze pilot performance in pointing the aircraft (which may be most sensitive to display clutter manipulations).

Mike Monroe did not think recording pilot decisions and decision times, as well as verbal protocols, in response to videos would be sufficient to detect the implications of clutter on performance, nor would the simulation be realistic enough as a basis for developing a measurement tool to support future display design and evaluation.

(Lance suggested using a cognitive walkthrough approach.

The cognitive walkthrough involves an analyst leading a system designer through an interface to identify errors a user might make. This usability evaluation method differs from user verbal protocols and is not applicable to the present experiment.)

Using simulators -

The potential use of X-plane was again mentioned, but concerns were raised regarding fidelity of the simulation, integration of the E/SVS imagery and generalizability of results.

Randy Bailey said that NASA could afford to provide access to the IFD for the Year 2 experiment.

He said that the simulator staff would support the project in developing the SVS wireframe models and the TCAS symbology.

Mike Monroe also supported the use of the IFD simulator in order to present more realistic conditions and to collect a battery of flight control performance measures to assess the implications of HUD clutter on pilot behavior.

Lance said that we could also assess flight technical error.

(This was done in the previous GA field studies with the SVS displays).

Mike Monroe suggested presenting nominal and high workload conditions in the IFD simulator, including normal turbulence and possibly an engine "burn-out".

He said this would be more representative of the range of realistic workload conditions and would provide insight into the role of clutter in nominal and critical flight circumstances.

There was a question about how we would get access to 10-20 pilots.

Randy said they would be recruited through Langley or the pilots from Dryden could fly in for the tests.

It was agreed that the IFD simulator would be used for the Year 2 experiment.

Current plan is to run experiment at Langley in May

Next steps:

The project team is to plan the experiment and develop all measurement instruments between now and the close of next February.

A pilot study will be conducted during spring break of 2008 (March 3-7)

Test conditions and procedures for admin. of experiment will be verified.

(Lance to check availability of testing in March, running in May – within the IFD)

Full experiment will be conducted at close of spring term (May 15-June 1)

Test with 10-20 pilots at Langley using IFD

(Lance says probably 12 commercial non-test pilots with HUD experience.)
(Must secure clearance for Sang-Hwan Kim in advance of this activity.)

Reporting requirements for Year 2:
Executive summary at end of July.
Final technical report due at beginning of October
Research briefing in October.

Anticipated Y2 Output:

- *Objective model of perceived clutter*
- *Subjective clutter preference threshold*
- *Objective performance clutter threshold*
- *Subjective clutter rating scale*
- *SPIE conference presentation (Lance suggests SPIE as a good venue for this work)*
- *Journal article*

Anticipated Y3 Activities:

- *Validate the model, thresholds, and subjective rating scales*
- *Extend to other displays to show generalizability and context-independence*
- *Beyond Y3 efforts may include design tool development to automate the process of objectively measuring clutter across displays*